DATASET 1

1) Explique o que você entendeu de cada tabela e a relação entre elas.

2) Escreva uma query SQL para encontrar os três produtos mais vendidos em termos de quantidade na tabela sales.

3) Escreva uma query SQL para calcular o lucro total por região.

4) Descreva como você implementaria um processo de ETL para carregar dados de um sistema legado para um data warehouse moderno.

Em Python:

5) Quais análises devem ser feitas nos dados fornecidos para pré-processamento? Descreva cada passo e demonstre em Python.

6) Calcule o total de vendas (SalesAmount) e o lucro por ano e por região.

7) Adicione uma coluna na tabela Sales com a classificação dos clientes em segmentos de alta, média e baixa lucratividade com base no lucro total gerado.

8) Qual seria o impacto esperado no lucro se o desconto médio fosse reduzido em 10%? Explique sua abordagem e resultados.

9) Utilizando modelagem preditiva, qual será a lucratividade futura das vendas.

DATASET 2

10) Construa duas DAG Airflow para uma data pipeline com as seguintes tarefas:

(1) Acesso a uma instância remota
(2) Extração do banco de dados da instância remota;
(3) Transformação dos dados localmente;

(4) Envio de relatórios. Configure o envio de um alerta caso a execução não seja executada corretamente duas vezes, com um intervalo de 5 minutos entre as execuções. As DAGs devem ser separadas em parte remota e local, sendo todas as etapas baseadas em eventos (com inicialização da segunda DAG baseada no término da primeira). Considere que a extração de banco de dados e a transformação de dados já estão prontos, nos seguintes arquivos, extract.py, transform.py e send.py e suas respectivas dependências req1.txt, req2.txt, req3.txt.

11) Foi-lhe solicitado para elaborar um ranking com as 20 maiores variações diárias de novos casos de covid. Neste ranking deve constar três colunas:

(1) a data dessa variação;
(2) a lista de cidades distintas;
(3) variação dos casos;
(4) o arquivo code.ipynb cumpre esses requisitos da melhor forma possível? Caso negativo, apresente o novo código.

12) Faça uma análise exploratória sobre os dados de covid e apresenta gráficos representativos e insights relevantes.

13) Em um projeto de treinamento de modelos de deep learning, a parte da anotação de dados é crucial para garantir um bom desempenho do modelo. Você está em um projeto de visão computacional cujas as imagens anotadas são segmentações que estão sobre determinadas peças, por exemplo: parabrisa. Quais estratégias você utilizaria para minimizar possíveis erros de anotações?

Instruções:

- Execute de preferência no Google Colab.
- Para cada questão, mostre os códigos e as respostas em seguida.
- Para questões teóricas, responda no notebook mesmo.
- Documente o processo de análise, incluindo todas as etapas de pré-processamento, análise e modelagem.